

# THE NATIONAL CONFERENCE ON NEXT GENERATION ASSESSMENT SYSTEMS

## HIGH-LEVEL MODEL FOR AN ASSESSMENT OF COMMON STANDARDS

This policy brief is based on a paper presented by Stephen Lazer, Educational Testing Service, at the National Conference on Next Generation K—12 Assessment Systems, March 2010. Download a copy of the final paper by Lazer, as well as other papers presented at the conference, at <http://www.k12center.org/publications.html>.

*This model proposes an integrated assessment system, not a single test, and focuses on the technical details of a summative system for elementary and middle schools. Using the Common Core Standards, it calls for end-of-year tests for Grades 3–8, which could be used to measure student growth if the standards cohere across grade levels. The summative/accountability components of the integrated assessment system might also include periodic classroom tests and collections of student work, which will be easier to implement if the common standards lead to a common sequence of learning objectives. Summative tests also could be used to provide information to subsequent diagnostic and formative assessments, particularly for students performing above or below grade level. The end-of-year and periodic components of the assessment should be computer-based. An assessment system for high schools would contain some of the same elements as the K–8 system. However, rather than choosing a specific approach, the paper offers policymakers two models: end-of-domain assessment or end-of-course assessment.*

Educators generally agree on the need for improved assessment systems, but there is far less consensus on the priorities for uses of the new assessments. It is impossible for one assessment or even one assessment system to fulfill everyone's goals, but there are some goals that are shared by various stakeholders. Other goals will require choices; an assessment system cannot do everything equally well.

The author of this paper made certain assumptions in creating this model. Among these are:

- The Common Core Standards will be adopted by a large number of states within 1–2 years.
- The Common Core Standards will cohere across grades so that assessments of the standards will support meaningful estimates of student growth.
- States and/or consortia may want to measure some attributes beyond those covered in the Common Core Standards.
- Universal computer-based testing will be possible in 3–4 years. However, the technology may not be adequate to test all students in mass administrations.



Center for  
K–12 Assessment  
& Performance Management

*Created by Educational Testing Service (ETS) to forward a larger social mission, the Center for K–12 Assessment & Performance Management has been given the directive to serve as an independent catalyst and resource for the improvement of measurement and data systems to enhance student achievement.*

- Major elements of the new assessment system must be made operational within 3–4 years, but the system can and will continue to evolve.
- Efficiencies from pooled test development and psychometric work will make modest increases possible in the per pupil operational costs of assessment, allowing some use of human scoring in the system.
- The goals of the common core assessment can only be met by an integrated assessment system, not by a single test.

Another assumption is that any assessment system must meet two overarching goals: new instructionally relevant measurement based on common standards and sound measurement that meets professional technical standards for high-stakes use.

## **Elements of a Comprehensive K–8 Assessment System**

The design calls for an integrated system with formative components and summative/accountability end-of-year assessments, which optimally may also include other components, such as periodic classroom tests and project-based components. Formative and summative components will work better if they measure the same standards. Moreover, an integrated system that includes formative assessments relieves pressure on summative assessments to provide varied information, especially for classroom-based decisions. An integrated system is versatile and could include interim tests, diagnostic tests, and item banks.

## **Structure of the Summative System**

The summative/accountability assessment system will include (but may not be limited to) end-of-year assessments at Grades 3–8 in English language arts (ELA) and mathematics. This will produce individual student scores, as well as aggregate scores. Annual testing is recommended to support student growth modeling and to take a snapshot of system progress at a fixed point in time. This would aid in comparability, a major goal of the system. For these reasons, end-of-year testing should be kept, although that does not preclude other data in an accountability system. The paper also discusses the possibility of summative/accountability tests distributed over the course of a year, which can be implemented in places where there is some degree of curricular uniformity.

The end-of-year tests may have at least two major components: common tests of the Common Core Standards and tests of state-specific content or augmentation (the Common Core agreement allows states to augment the Standards with up to 15% state-specific standards). The model allows states to pass up state-customized components, if they wish.

It may be premature to discuss item types before the standards are established, but enough is known about the emerging standards to make some general points. The construct and measurement needs of the system will require a range of exercise types, from selected response to short-answers and more extended tasks. This mix of item types is likely because of the college-ready expectations that require information on students' abilities in a variety of areas, such as problem-solving and conducting critical

analyses. The items and tests should be developed with an awareness of how students learn. Teaching to the test would be less of a problem if the test reflects learning progressions and models good learning and instruction, while maintaining technical quality.

According to the paper, several issues will need to be addressed during the design phase. Among these questions are how to deal with the use of audiovisual resources and interactive tasks, how to address the standards for speaking and listening in a summative assessment, the size of item pool needed to ensure security, and the length of individual tests at different grade levels.

## **Computer-Based Assessments**

A major question for test designers is how much technology and how soon? This paper calls for aggressive use of technology in the testing program and argues that standardized assessment components of the system should be computer-based, with traditional paper/pen reserved for special accommodations. There are several reasons for this, including:

- Emerging standards in ELA and mathematics—and eventually in science—likely will define constructs that only can be measured through the use of technology.
- Technology allows for the use of a range of forward-looking item types such as digital content and formats.
- Testing some skills such as writing on paper may yield invalid results because students are accustomed to doing their work on computers.
- Technology allows for flexible (adaptive) testing and electronic scoring of some items, which will broaden the range of items.
- Technology facilitates more effective dissemination of student responses to teachers and the use of assessment development and scoring for professional development.
- Technology speeds up access to results; makes a broad range of accommodations possible; and if it is the only delivery tool, simplifies issues with comparability.

The summative assessment system and end-of-year tests in particular should make use of adaptive testing administration for several reasons. It allows shorter testing times than linear testing and the use of assessment pools that cover more rigorous standards. It could identify standards that are particularly difficult for students, while also allowing a bigger bang for the buck from open ended/performance-based testing. It also can accommodate extended windows for testing while maintaining high security.

In an adaptive system, the use of items that require human scoring could represent challenges. But there are ways to address them, such as multi-stage testing with machine-scored testing followed later by items requiring human scoring.

The new end-of-year assessment system must be innovative, but at the same time affordable, sustainable, and able to provide rapid scores. Expanded items types are needed that push the limits on

what can be scored electronically. Scaled scores and status indicators from an assessment remain essential, but technology must allow us to develop better ways of analyzing data from assessments (e.g., the steps students take in writing essays or engaging in mathematics simulations).

The paper argues that even given advances in electronic scoring, if the assessment is to measure key outcomes human scoring will be necessary for some items, even though it raises issues about affordability. The added value of human scoring is the professional development payoff for teachers who do the scoring, provided the system uses the professional development for maximum impact while minimizing the burden.

## **Measuring Student Progress and Reporting Status Indicators**

Given the interest in student growth measures, the assessment system should support cross-grade comparability of scores. Various cross-grade comparability models can be used in the Grade 3–8 elements of the system. Measuring growth at the high-school level may be more problematic, depending on the high-school model chosen.

If status (proficiency scores) are to be used, it would make more sense from a measurement perspective to apply them to the summative system as a whole and not to specific standards. The content standards must give sufficient guidance on acceptable levels of performance.

## **Using Periodic Assessments and Project-Based Scores in a Summative System**

Combined with end-of-year testing, periodic testing should provide much richer information on what students know and can do and should have instructional value. This could be accomplished through single or multiple standardized assessments over the course of a year. Another model is to use data from standardized projects conducted over the course of study (e.g., research papers, lab reports) instead of or in addition to periodic assessments. This would encourage the use of tasks usually omitted from large-scale testing but reflective of good instruction. To the extent to which this would rely on local choice, it raises issues of comparability over time and across jurisdictions, but the problems with both kinds of assessments can be eased if the common standards lead to a common sequence of learning objectives. These components should be initially used with low stakes and then, when ready, phased into the summative/accountability system.

## **Addressing High-School Assessment**

The major elements of the K–8 assessment system should be applicable to the high school assessment system—an array of item types, a combination of automated and human scoring and delivery through technology. Policymakers, however, have at least two models to choose from: end-of-domain assessment or end-of-course assessment.

End-of-domain assessments would be given at the point considered right to judge college/career readiness skills mastery (Grades 11–12) for ELA and mathematics. These tests would be adaptive to allow them both to cover rigorous content and skills and to report meaningful results for all students.

This kind of testing allows for comparisons across schools. If students are testing at different times in their high-school careers (for example, upon entry into high school and then in Grade 11), this model could be used to measure student growth. Rapid reporting of scores would not be imperative, and local educators would have maximum flexibility in curriculum decisions (end-of-course models will constrain some choices). On the other hand, because end-of-domain assessments are not linked to specific courses, they will probably be less effective in providing feedback to teachers or serving as indicators in teacher accountability systems.

The end-of-course assessment model forges direct links between assessment and instruction and could drive instructional improvement. It would include end-of-course tests and could include periodic and project-based assessments and be delivered by computer (if scores are available quickly, they could become part of a course grade). A drawback is that while data are comparable at the student level, aggregation to the school level and higher becomes potentially problematic, given different course-taking patterns. End-of-course assessments also create certain issues in measuring student growth data.

Deciding between these high-school models will depend on the goals for assessment, and it might be possible to use both models in an assessment system, although this would be expensive. In any case, it is important to have research and validity evidence that supports the intended use of the scores from the assessment. Even if international benchmarks set the base, these benchmarks need to be updated continually, which will change test specifications. The claim that the tests measure college readiness needs to be validated, and gathering such data should be part of the design from the beginning.

## For More Information

For more information on this assessment system model, please see the paper by Stephen Lazer:

Lazer, S. (2010). *High-level model for an assessment of common standards*. Retrieved from <http://www.k12center.org/publications.html>.

For more information on the National Conference on Next Generation Assessment Systems, please see: <http://www.k12center.org/events.html>.